DOCUMENT RESUME

ED 255 875 CS 007 989

AUTHOR Alvermann, Donna E.; And Others
TITLE Assessment of Classroom Interaction Dynamics.
PUB DATE 84
NOTE 12p.; Paper presented at the Annual Meeting of the National Reading Conference (34th, St. Petersburg, FL, November 28-December 1, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Communication; Classroom Observation Techniques; Classroom Research; Comparative Analysis; Discussion (Teaching Technique); *Interaction; Interaction Process Analysis; *Interrater Reliability; Judges; *Measurement Techniques; Teacher Evaluation; *Teacher Student Relationship; Teaching Methods; *Test Reliability
IDENTIFIERS *Assessment of Classroom Interaction Dynamics

ABSTRACT

To help teachers develop an awareness of how they structure a discussion, an instrument was constructed called the Assessment of Classroom Interaction Dynamics (ACID). Two expert judges and 26 trainees then participated in a study (1) to estimate interrater reliability between expert judges in the use of the ACID, (2) to assess the validity of the judgments made by the trainees, and (3) to estimate the reliability of the categories for detecting teacher differences. Data analyses revealed mixed results. On the one hand, the relatively good level of interrater reliability involving the expert judges indicated an observational instrument capable of yielding results not critically dependent upon the identity of a particular rater. Also, the relatively high level of agreement between the trainees and both experts suggests the instrument was easy to learn to use. On the other hand, ACID's reliability in differentiating among teachers was disappointing. None of the estimated categorical reliabilities reached an acceptable level. A copy of the assessment instrument is attached. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Donna E. Alvermann
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DONNA E. ALVERMANN
JOSEPH WISENBAKER
DEBORAH R. DILLON
University of Georgia

ED255875

Assessment of Classroom Interaction Dynamics

Nationwide assessments of reading achievement in American schools consistently reveal that very few students learn to interpret critically what they read. A report issued by the last National Assessment of Educational Progress (NAEP, 1981) stated that "...while students learn to read a wide range of material, they develop very few skills for examining the nature of the ideas that they take away from their reading" (p. 2). The report went on to suggest that this finding may be linked to patterns of teacher-student interaction which foster recitation, not discussion. For a true discussion to occur, according to Roth, Smith, and Anderson (1984), "Teachers need to know more than just what questions to ask; they also need to know how to respond to student statements" (p. 287). Implicit in Roth et al.'s recommendation is the notion that teachers must develop an awareness of how they

CS007989

structure a discussion; for example, how they allot turntaking and what types of responses sustain or curtail a discussion.

To help teachers develop this awareness, we constructed an instrument called the Assessment of Classroom Interaction Dynamics (ACID) (Table 1). It was designed for supervisory use as well as self-assessment. The categories and properties of ACID evolved from a year-and-a-half long study of 24 middle school classrooms (Alvermann, O'Brien, Dillon, & Smith, 1984). Briefly, that study described the nature of all verbal interaction patterns which occurred during each teacher's videotaped class discussions of content area reading assignments. Data obtained from the transcriptions of the videotapes and from the accompanying field notes were simultaneously analyzed and reduced into categories and properties using the constant comparative methodology of Glaser and Strauss (1967).

The purpose of the present study was threefold: to estimate interrater reliability between expert judges in the use of the ACID instrument, to assess the validity of the judgments made by the trainees, and to estimate the reliability of the categories for detecting teacher differences.

Method

Subjects

Two expert judges and 26 trainees participated in this study. The experts were so named because they had been members in the original group that developed the categories and properties of the ACID instrument. With the exception of one individual, the trainees were all enrolled in one of two graduate level reading education courses. Seventeen were masters level students in a general reading methods course and 8 were doctoral level students in a research seminar. The

instructor for that seminar also participated as a trainee. The majority of the participants were full-time teachers.

Materials

In addition to the ACID instrument, materials used in this study included a practice videotape, three experimental videotapes, a handout defining the categories and properties of ACID, and overhead transparencies of that handout. The practice and experimental videotapes each consisted of three 5-minute segments that represented the beginning, middle, and end of a teacher's entire class discussion.

Procedures

A series of preliminary administrations of the ACID instrument were conducted to determine an appropriate set of procedures, scoring schemes, and time requirements. For standardization purposes, the two expert judges followed a typed set of instructions during each of the three 45-minute training sessions held at one week intervals. A brief description of the training sessions follows.*

The first session began with a short history of the research that led to the construction of the ACID instrument. The objectives of the training sessions were explained (i.e., to obtain reliability and validity estimates of the ACID instrument), and the six categories and their respective properties were defined. Further clarification of the categories and properties was achieved by displaying transparencies that contained transcribed segments of actual classroom discussion. For homework, the trainees were instructed to study the various category and

*For a complete set of procedures and materials, write to the first author at 309 Aderhold Building, University of Georgia, Athens, GA 30602.

property definitions of the ACID instrument so that they could use them with ease during the next session.

Session two began with a brief review of the categories and properties. The trainees were then shown how to code the ACID instrument. They were directed to place a check mark next to the appropriate category each time they observed one of the videotaped teachers behaving in a manner that reflected one of the category's properties. After each of the three 5-minute segments, the practice videotape was stopped and a discussion was held so that the trainees could compare their codings with those of the expert judge. For homework, the trainees were again directed to study the various categories and their properties so that they could use them in coding the experimental videotapes one week later.

Session three also began with a review of the categories and properties. Next, the three experimental videotapes (each with a 5-minute beginning, middle, and ending segment) were shown to the trainees who were instructed to check the categories observed during the various teacher-student interactions. Unlike session two, however, discussion was not permitted after any of the segments. At the conclusion of the last videotape, the trainees were instructed to distribute 60 points across the 6 categories. They were told that the number of checks in each category did not necessarily dictate the point distribution; rather, they could distribute the points according to the saliency of a particular category in relation to all the others.

Analyses Performed

The data provided by the subjects were analyzed in several different ways to address the primary questions posed in the study. Interrater reliability in the use of the scales between expert judges

and the validity of the judgments made by the trainees were estimated. Additionally, the reliability of each category for detecting teacher differences was estimated for the expert judges and for a subset of trainees.

Estimating the interrater reliability of the expert judges involved using multifactor reliability tests as outlined by Nunnally (1982). This approach treated the judgments of the experts as a dependent variable in a four factor design involving subjects, raters, occasions, and categories. A four-way analysis of variance was carried out with variance components estimated and each factor treated as random. In the estimation of interrater reliability, all variance components involving raters were treated as error variance.

The validity of the judgments made by the trainees was assessed by pairing each of the sets of a trainee's judgments with those of the senior expert judge and computing a correlation between the two sets of judgments. All trainees for whom this correlation was not statistically significantly lower than that between the two expert judges had their data retained in the subsequent analyses.

The reliability of the instrument for detecting differences among teachers was estimated separately for the expert judges and for the trainees remaining from the previous stage of the analysis. This was done for each of the six individual categories of interactions addressed by the observational instrument. The methodology employed here was similar to that used to assess interrater reliability among the expert judges. For each category the observations served as the dependent variable in a three-way analysis of variance based on the design factors of rater, occasion, and subject. Treating each of these design factors as random, the variance components were estimated. These were then used

to estimate the reliability of each category for detecting differences among subjects under four different conditions. The conditions were defined by the number of judges (1 or 2) crossed with the number of observational periods (3 or 10).

All statistical tests of significance were carried out using an alpha level of .05. The computations were performed on an IBM 3081 using the Statistical Analysis System (SAS, 1982).

Results

Interrater Reliability Between Expert Judges

Since the intent of this analysis was to determine the interrater reliability of the expert judges, all variance components involving "raters" (including the "error" component) were treated as error variance. The remainder were treated as true variance. The estimated variance of the true score terms summed to 6.3833 while the estimated variance of the error score terms summed to 2.2934. In each instance, negative values were disregarded. The relation of estimated true score variance to estimated total variance was .74.

Validity of Trainee Judgments

Overall, quite a few of the trainees provided judgments that correlated well with those of the senior expert. All of the correlations were statistically significantly larger than zero (alpha = .05). Additionally, the judgments made by 16 of the trainees correlated with those of the senior expert no more poorly than with those of the other expert (alpha = .05 for each test).

Reliability of Categories for Detecting Teacher Differences

The previous analysis established the extent to which there was interrater reliability among the expert judges in the use of the categories and the extent to which the judgments of the trainees agreed

with those of the senior expert. In a very real sense, having adequate levels of each of these is a necessary condition for the future application of the instrument. This section of the paper focuses on the extent to which the instrument is capable of reliably detecting differences among teachers on each of the categories.

For each category, reliability indices associated with the detection of differences among teachers were estimated for the expert judges and for the trainees who had the better agreement with the senior judge. This was done using the generalizability approach outlined by Nunnally (1982). In each instance, reliability was estimated under the four conditions defined by crossing the number of raters (1 or 2) with the number of observational periods (3 or 10). Only the variance term associated with "subjects" was treated as true score variance in this set of analyses.

Category reliability based on expert judges. The first stage in the estimation of the reliability with which expert judges can make use of each of the six categories involved performing analysis of variance. The category frequencies served as the dependent variable while the factors in the design consisted of raters, subjects, and occasions. The variance components for each category were then combined to yield reliability estimates. As mentioned earlier, the estimated variance among subjects was the only true score variance term in these analyses. The error terms were combined by weighting each by factors determined from the source of the error term and the condition for which it was to serve as an estimate. The estimated reliabilities based on the expert judges were very small. Of the reliability estimates which could be compared, none were based on a statistically significant between subjects variance (alpha = .05).

Category reliability based on "better" trainees. Reliability estimates for the trainees were computed making use of the same procedures as with the expert judges in the previous section. Again, none of the estimated categorical reliabilites reached an acceptable level.

Summary and Conclusions

Overall, the analyses reported in this paper represent a mixed set of results. On the one hand, the relatively good level of interrater reliability involving the expert judges is indicative of an observational instrument capable of yielding results that are not critically dependent upon the identity of the particular rater involved. That two-thirds of the trainees provided judgments which correlated nearly as well with those of the senior expert as with those of the other expert offers a reasonable expectation that this instrument is relatively easy to train others to use.

On the other hand, the reliability with which teachers could be differentiated in their characteristics on the individual categories is clearly disappointing. Further research in this area will focus on whether checking specific properties within the categories (as opposed to checking only the categories) will yield better results.

References

Alvermann, D. E., O'Brien, D. G., Dillon, D. R., & Smith, L. C. (1984, May). Textbook reading assignments: An analysis of teacher-student discussion. Paper presented at the Center for the Study of Reading's Fifth Annual Conference on Reading Research, Atlanta, GA.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. New York: Aldine.

National Assessment of Educational Progress (1981). Reading, thinking, and writing: Results from the 1979-80 national assessment of reading and literature (Report No. 11-L-01). Denver, CO: National Assessment of Educational Progress.

Nunnally, J. C. (1982). Reliability of measurement. In E. Mitzel (Ed.), Encyclopedia of educational research (5th ed.) (pp. 1585-1601). New York: Macmillan.

Roth, K. J., Smith, E. L., & Anderson, C. W. (1984). Verbal patterns of teachers: Comprehension instruction in the content areas. In G. Duffy, L. Roehler, & J. Mason (Eds.). Comprehension instruction: Perspectives and suggestions. Longman.

SAS Institute. (1982). SAS user's guide: Statistics. Cary, NC: SAS Institute, Inc.

THE ACID TEST (ASSESSMENT OF CLASSROOM INTERACTION DYNAMICS)

CATEGORIES/PROPERTIES	BEGINNING SEGMENT	MIDDLE SEGMENT	END SEGMENT	POINTS
CONTROL: *maintenance of order *shifts in control *physical space *puts words in mouth- asks for agreement				
SUSTAINING DISCUSSION: *critical questioning *intonation *clarifying *reinforcement/evaluation *ambiguous ok, all right, uh-huh *response paraphrased/explained *vocabulary discussed *calls on--doesn't give up				
SENSE OF AUDIENCE: *concern with balance *encourages participation *teacher admits-doesn't know *teacher lectures *permissive behavior by teacher *teacher misreads student *teacher in playful mood *student-student interaction				
PACING: *rephrase to clarify *can-you-guess-what-I'm-thinking game *patterned rhythm to questions/answers				
USE OF TEXTUAL MATERIALS: *text to verify *text not open *indirect reference to text *use text when answers are not forthcoming *use text to refocus *text basis for paraphrasing responses				
RELEVANCE OF CONTENT: *teacher draws analogies/ hypothetical situations *inaccurate information given by teacher *personal anecdotes related *teacher refers to newspapers/ radio/ television to relate				
			TOTAL POINTS:	*60

BEST COPY AVAILABLE

*COMMENTS (opt.)